NLS (Nuclear Localization Signal) Prediction (MOLECULAR BIOLOGY AND INFORMATION?Biological Information Science)

NLS (Nuclear Localization Signal) Prediction

Keun-Joon Park and Minoru Kanehisa

The nucleus of eukaryotic cells contains many nuclear proteins that function for delivery of molecular information between cytosol and nucleus, and for control of gene expression. After the synthesis on the ribosomes in the cytoplasm, these proteins enter the nucleus through pore complexes in the nuclear envelope. Nuclear proteins are transported into the nucleus, if they contain nuclear localization signals (NLSs). In this work we developed a method that predicts a location of NLS on a query protein sequence by computational analysis. We employed Hidden Markov Model (HMM) in our method to find NLSs in the amino acid sequences. The prediction performance was assessed by leave-one-out cross-validation.

keywords: Nuclear transport / Protein sorting / Database / Bioinformatics

In eukaryotic cells, there are functionally distinct, membrane-bounded compartments. The intracellular compartments in eukaryotic cells contain their own characteristic proteins with different functions. Most nuclear proteins move into nucleus through the nuclear pores that penetrate nuclear envelope (Fig. 1). Nuclear pore is formed by a large, complex structure known as the nuclear pore complex (NPC). The selective transport of proteins through the NPC is performed by their own nuclear localization signals (NLSs). The first NLS was found from SV40 T antigen as a short cluster of five contiguous positively charged residues in the sequence ^{126}PKKKRKV^{132} (3). A family of simple NLSs of this type were generally characterized by one short basic stretch of sequence (4-8 residues) containing several lysine and arginine residues. The precise location of an NLS within the amino acid sequence of a nuclear protein is not important unlike other signal peptides (4). Robbins et al. found other type of NLS in the nucleoplasmin, the major nuclear protein of the xenopus oocyte (5,6). This type of signal, known as bipartite NLS motif, contains two interdependent positively charged clusters separated by a mutation tolerant linker region of 10-12 amino acids.

In this study we developed a method that predicts the location of NLS and the possibility of a nuclear protein by computational analysis of NLSs. For this purpose, we constructed data sets from the SWISS-PROT protein sequence database (simple NLS 100 entries and bipartite

MOLECULAR BIOLOGY AND INFORMATION — Biological Information Science —

Scope of research

This laboratory aims at developing theoretical frameworks for understanding the information flow in biological systems in terms of genes, gene products, other biomolecules, and their interactions. Toward that end a new deductive database is being organized for known molecular and genetic pathways in living organisms, and computational technologies are being developed for retrieval, inference and analysis. Other studies include: functional and structural prediction of proteins from sequence information and development of sequnece analysis tools.

Prof
KANEHISA, Minoru
(D Sc)

Instr
GOTO, Susumu
(D Eng)

Instr
OGATA, Hiroyuki
(D Sc)

Students:
SUZUKI, Kenji (DC)
KIHARA, Daisuke (DC)
KAWASHIMA, Shuichi (DC)
PARK, Keun-Joon (DC)
HATTORI, Masahiro (DC)
BONO, Hidemasa (DC)
IGARASHI, Yoshinobu (DC)
KATAYAMA, Toshiaki (DC)
SATO, Kiminobu (DC)
TAKAZAWA, Fumi (MC)
TANIGUCHI, Takeaki (MC)
NAKAO, Mitsuteru (MC)
OKUJI,. Yoshinori (MC)
LIMVIPHUVADH, Vachiranee (RS)

Research Fellow:
SATO, Kazushige (RF)

NLS 75 entries). We used Hidden Markov Model (HMM) for predicting NLSs within the amino acid sequences and for calculating the strength of the signals. We trained two (simple and bipartite) HMMs for unaligned sequences of NLSs in our data sets.

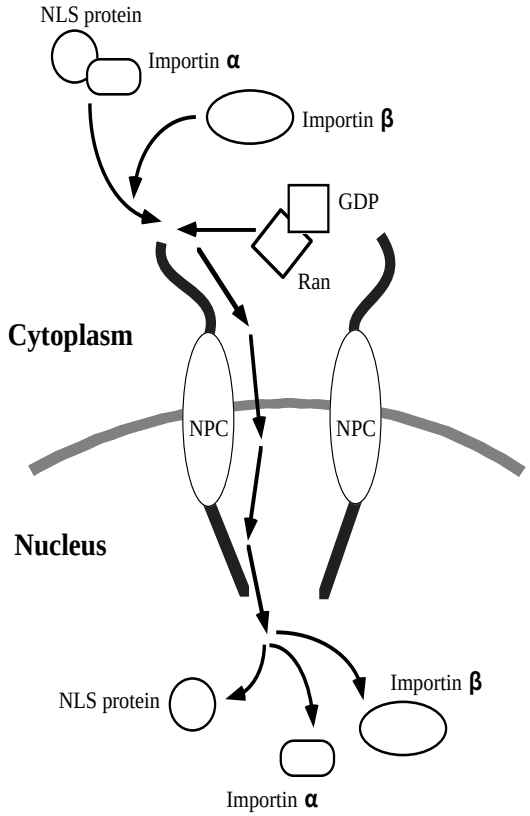

Figure 1. Nuclear protein import
(NLS: Nuclear Localization Signal; NPC: Nuclear Pore Complex)

The performance of the HMMs to recognize NLSs trained on the data sets may be improved by using calculation with fewer different kinds of amino acids (alphabet characters). In this work we changed all arginines (R) to lysines (K) in the sequences of the data sets and the query proteins, since these two positively charged residues would likely to have the same function in NLSs. It seems that the NLSs are usually situated at the extended loop structure that is accessible to the solvent (7). Therefore we examined the frequency of amino acid residues in the NLSs and their secondary structure propensity in order to reduce the number of amino acid types for training. Some amino acids were selected for HMM training and the other amino acid residues were changed into residue Xs in both of the two NLS cases.

To test the performance of the HMM, each protein sequence in the data set was selected once as a test sequence for the HMMs trained for the other sequence in the data set. Specifically, the performance was assessed by leave-one-out cross-validation ; the leave-one-out idea is often called "jack-knife test". In the training step, one NLS is left out from the training set. Then the protein sequence containing the NLS is used for checking the performance of the trained HMM. The HMM for simple NLSs was trained for 99 from 100 samples and tested on the remaining one. This process was repeated 100 times for different test sequence. The second HMM for bipartite NLSs was trained for 74 from 75 data set entries and yielded 75 different HMMs. We examined whether two HMMs can recognize their own NLS sequences in their data set or not. As a result, the prediction accuracy of our method was 88.0% and 90.7% for simple and bipartite NLSs, respectively. In conclusion, we confirmed that the two HMMs could predict NLS in amino acid sequences with high performance.

Acknowledgments

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas ‘Genome Science’ from the Ministry of Education, Science, Sports and Culture in Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

1. D. Kalderon, W.D. Richardson, A.F. Markham and A.E. Smith, Nature, **311**, 33 - 38 (1984).
2. J. Garcia-Bustos, J. Heitman and M.N. Hall, Biochim. Biophys. Acta, **1071**, 83 - 101 (1991).
3. J. Robbins, S.M. Dilworth, R.A. Laskey, and C. Dingwall, Cell, **64**, 615 - 623 (1991).
4. C. Dingwall and R.A. Laskey, Trends Biochem. Sci., **16**, 478 - 481 (1991).
5. J. Löwe, D. Stock, B. Jap, P. Zwickl, W. Baumeister, R. Huber, Science, **268**, 533 - 539 (1995).